Matrix-Element

Calculate a given process (e.g. gluino pair)

- Determine the production mechanism

diagram 1   QCD=2, QED=0   diagram 2   QCD=2, QED=0
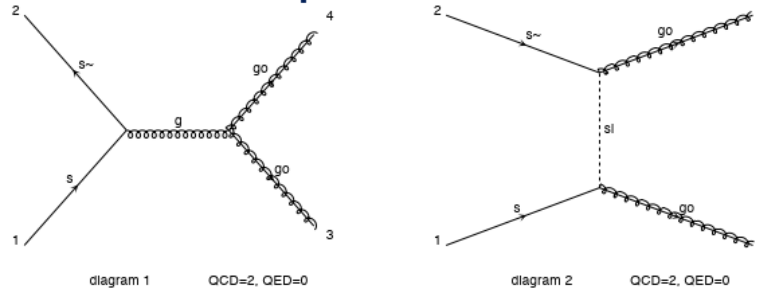
Easy enough

- Evaluate the matrix-element

$$|\mathcal{M}|^2$$ ➡Need Feynman Rules!

Hard

- Phase-Space Integration

$$\sigma = \frac{1}{2s} \int |\mathcal{M}|^2 d\Phi(n)$$

Very Hard
(in general)

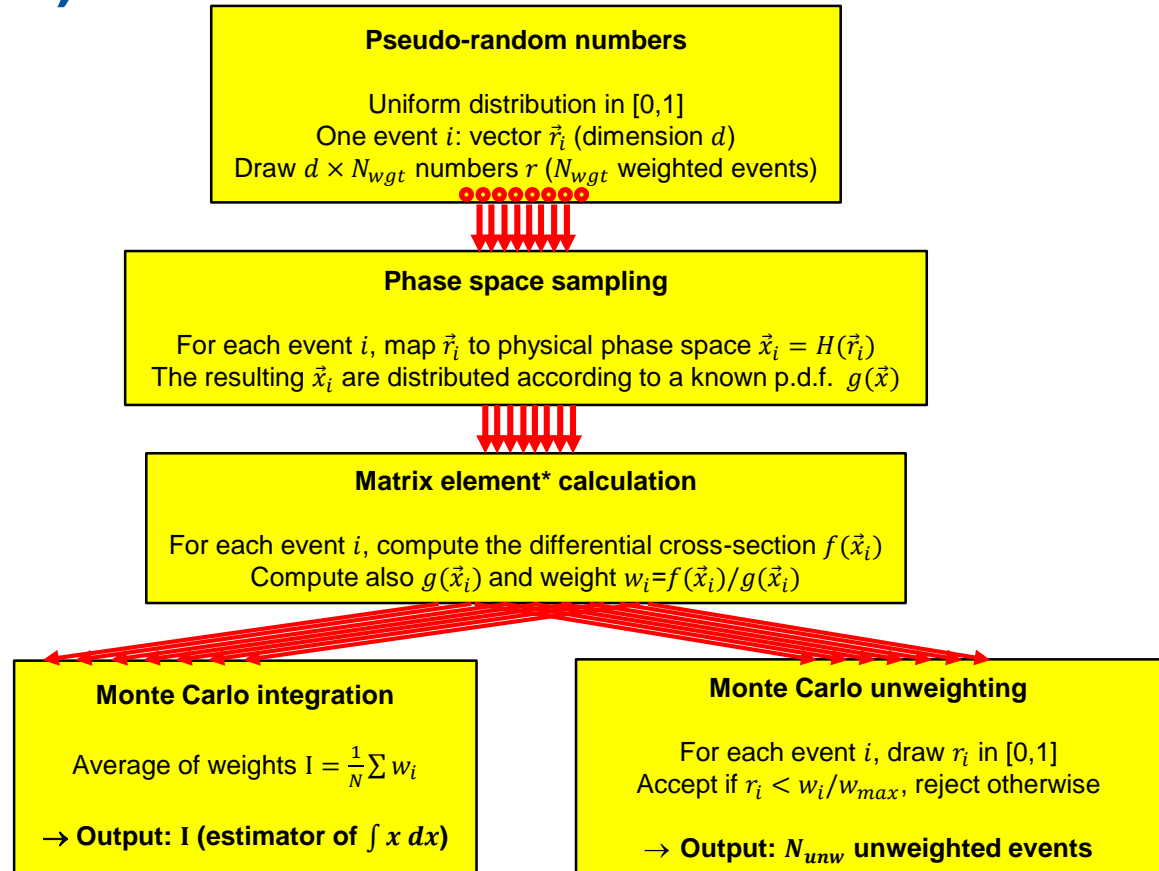Mattelaer Olivier   Milan 2020   2

MadGraph GPU development 22 June 2020 – O. Mattelaer
https://indico.cern.ch/event/932057/contributions/3917111

Team Madgraph   Sheffield GPU hackathon – 10 Jul 2020   1

MC generators: _simplified_ computational anatomy

Pseudo-random numbers

Uniform distribution in [0,1]
One event i: vector $\vec{r}_i$ (dimension d)
Draw $d \times N_{wgt}$ numbers r (N_{wgt} weighted events)

Phase space sampling

For each event i, map $\vec{r}_i$ to physical phase space $\vec{x}_i = H(\vec{r}_i)$
The resulting $\vec{x}_i$ are distributed according to a known p.d.f. $g(\vec{x})$

Matrix element calculation

For each event i, compute the differential cross-section $f(\vec{x}_i)$
Compute also $g(\vec{x}_i)$ and weight $w_i = f(\vec{x}_i)/g(\vec{x}_i)$

Phase space integration optimization
Compute w_{max}
Optimize phase space sampling $H(\vec{r}_i)$:
minimize variance of I,
_OR_ maximize unweighting efficiency

Monte Carlo integration

Average of weights $I = \frac{1}{N}\sum w_i$

→ **Output: I (estimator of $\int x\,dx$)**

Monte Carlo unweighting

For each event i, draw r_i in [0,1]
Accept if $r_i < w_i/w_{max}$, reject otherwise

→ **Output: N_{unw} unweighted events**

Parton showers (PS)

Hadronisation and decay

Particle-level filtering

Detector simulation

UCLouvain

# *Data-parallel paradigms (GPUs and vectorization)*

Generators should lend themselves naturally to **data-parallel paradigms**?
- **SPMD**: Single Program Multiple Data (GPU accelerators)
- **SIMD**: Single Instruction Multiple Data (CPU vectorization: AVX…)
- The computationally intensive part, the matrix element $f(\vec{x}_i)$, is *the same function* for all events i (in a given category of events)

- Unlike detector simulation (frequent if/then branches: on GPUs, branch divergence)

Pseudo-random numbers

Uniform distribution in [0,1]
One event i: vector $\vec{r}_i$ (dimension d)
Draw $d \times N_{wgt}$ numbers r (N_{wgt} weighted events)

Phase space sampling

For each event i, map $\vec{r}_i$ to physical phase space $\vec{x}_i = H(\vec{r}_i)$
The resulting $\vec{x}_i$ are distributed according to a known p.d.f. $g(\vec{x})$

Matrix element* calculation

For each event i, compute the differential cross-section $f(\vec{x}_i)$
Compute also $g(\vec{x}_i)$ and weight $w_i = f(\vec{x}_i)/g(\vec{x}_i)$

Monte Carlo integration

Average of weights $\mathrm{I} = \frac{1}{N}\sum w_i$

$\rightarrow$ **Output: I (estimator of $\int x\,dx$)**

Monte Carlo unweighting

For each event i, draw r_i in [0,1]
Accept if $r_i < w_i/w_{max}$, reject otherwise

$\rightarrow$ **Output: N_{unw} unweighted events**

*Note for software engineers: these calculations do involve some linear algebra, but "matrix element" does not refer to that! Here we compute one "matrix element" in the S-matrix (scattering matrix) for the transition from the initial state to the final state
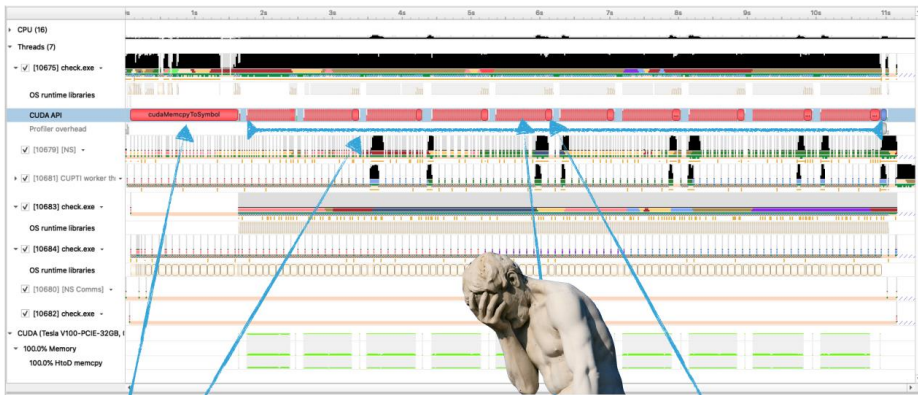
HOW MUCH CPU IS USED FOR THE MATRIX ELEMENT CALCULATION?
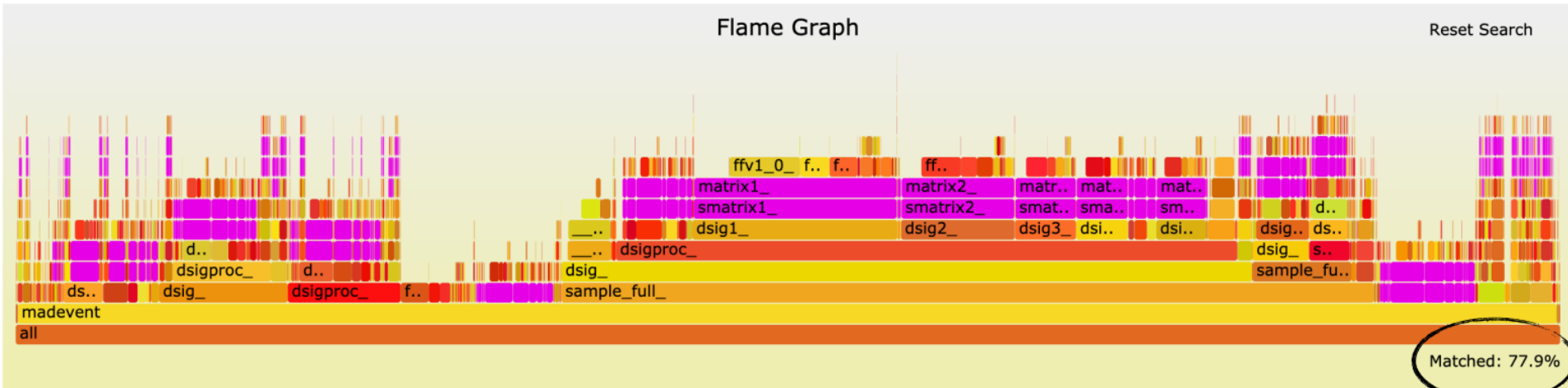
▶ E.g. real world CMS example: p p –> l+ l- j j j j / h @0

  ▶ Madgraph/MadEvent (Fortran), 10^5 events

CERN IT-SC group meeting 11 June 2020 - S. Roiser
https://indico.cern.ch/event/921329/contributions/3881034

Team Madgraph   Sheffield GPU hackathon – 10 Jul 2020   4

ONE PROFILING EXAMPLE WITH NSIGHT SYSTEM

▶ Copy parameters to the kernel

▶ Iterate 10 times through the kernel passing inputs, receiving outputs

ONE PROFILING EXAMPLE WITH NSIGHT SYSTEM — BETTER
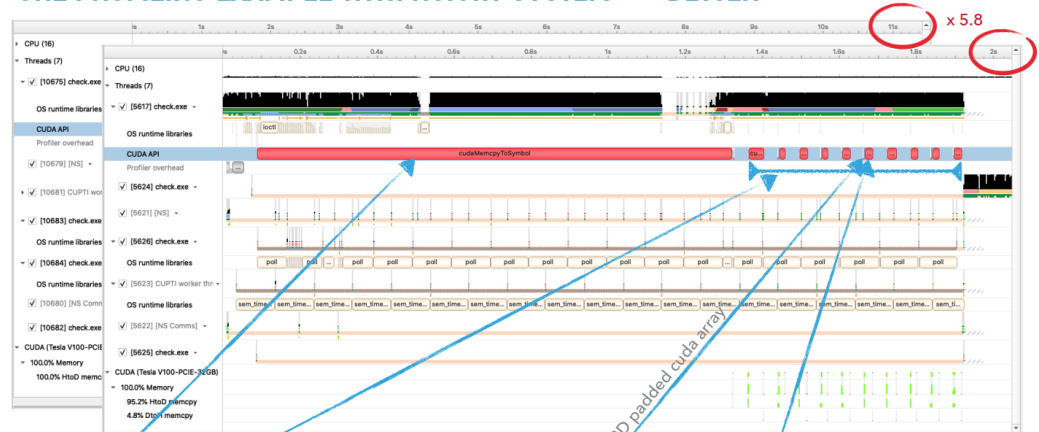
x 5.8

Pass 3D padded cuda array

▶ Copy parameters to the kernel

▶ Iterate 10 times through the kernel passing inputs, receiving outputs

events / second

cpu version	297311.8638
numevents	6291456

GPU Total Time in Wavefunctions

gpu	blocks/grid	1	2	4	8	16	32	64	128	256	--> threads / block
	1	1.95E+02	1.01E+02	4.96E+01	2.46E+01	1.30E+01	8.11E+00	5.06E+00	2.88E+00	2.07E+00	
	2	1.01E+02	4.91E+01	2.57E+01	1.31E+01	7.23E+00	5.09E+00	2.94E+00	1.85E+00	1.54E+00	
	4	5.06E+01	2.46E+01	1.30E+01	7.42E+00	5.09E+00	3.07E+00	1.88E+00	1.42E+00	1.22E+00	
	8	2.46E+01	1.29E+01	7.11E+00	5.10E+00	3.07E+00	1.84E+00	1.42E+00	1.14E+00	1.03E+00	
	16	1.30E+01	6.92E+00	5.08E+00	2.89E+00	1.84E+00	1.42E+00	1.15E+00	1.00E+00	9.44E-01	
	32	7.00E+00	5.09E+00	3.04E+00	1.84E+00	1.41E+00	1.16E+00	1.00E+00	9.27E-01	9.14E-01	
	64	5.07E+00	2.94E+00	1.83E+00	1.42E+00	1.16E+00	1.01E+00	9.34E-01	9.03E-01	8.02E-01	
	128	3.03E+00	1.84E+00	1.42E+00	1.16E+00	1.01E+00	9.36E-01	8.91E-01	7.98E-01	4.71E-01	
	256	1.84E+00	1.42E+00	1.16E+00	1.01E+00	9.39E-01	9.01E-01	7.91E-01	4.60E-01	2.95E-01	
	512	1.42E+00	1.17E+00	1.01E+00	9.31E-01	8.91E-01	7.88E-01	4.63E-01	2.82E-01	1.96E-01	
	1024	1.32E+00	1.08E+00	9.78E-01	9.14E-01	8.04E-01	4.64E-01	2.90E-01	2.02E-01	1.41E-01	
	2048	1.21E+00	1.03E+00	9.53E-01	8.17E-01	4.64E-01	2.87E-01	2.03E-01	1.38E-01	1.10E-01	

GPU Events/second

gpu		1	2	4	8	16	32	64	128	256
	1	3.23E+04	6.22E+04	1.27E+05	2.55E+05	4.82E+05	7.75E+05	1.24E+06	2.18E+06	3.04E+06
	2	6.21E+04	1.28E+05	2.45E+05	4.81E+05	8.70E+05	1.24E+06	2.14E+06	3.41E+06	4.10E+06
	4	1.24E+05	2.56E+05	4.83E+05	8.48E+05	1.24E+06	2.05E+06	3.34E+06	4.42E+06	5.17E+06
	8	2.56E+05	4.86E+05	8.85E+05	1.23E+06	2.05E+06	3.41E+06	4.42E+06	5.51E+06	6.09E+06
	16	4.83E+05	9.09E+05	1.24E+06	2.18E+06	3.41E+06	4.45E+06	5.45E+06	6.28E+06	6.67E+06
	32	8.99E+05	1.24E+06	2.07E+06	3.42E+06	4.47E+06	5.45E+06	6.28E+06	6.79E+06	6.89E+06
	64	1.24E+06	2.14E+06	3.43E+06	4.42E+06	5.41E+06	6.25E+06	6.73E+06	6.97E+06	7.85E+06
	128	2.07E+06	3.42E+06	4.42E+06	5.41E+06	6.25E+06	6.72E+06	7.06E+06	7.88E+06	1.34E+07
	256	3.42E+06	4.42E+06	5.41E+06	6.25E+06	6.70E+06	6.98E+06	7.95E+06	1.37E+07	2.13E+07
	512	4.42E+06	5.40E+06	6.22E+06	6.76E+06	7.06E+06	7.99E+06	1.36E+07	2.23E+07	3.20E+07
	1024	4.78E+06	5.81E+06	6.43E+06	6.88E+06	7.83E+06	1.36E+07	2.17E+07	3.11E+07	4.46E+07
	2048	5.19E+06	6.08E+06	6.60E+06	7.70E+06	1.36E+07	2.19E+07	3.10E+07	4.56E+07	5.71E+07

Factor GPU/CPU

gpu		1	2	4	8	16	32	64	128	256
	1	0.11	0.21	0.43	0.86	1.62	2.61	4.18	7.34	10.24
	2	0.21	0.43	0.82	1.62	2.93	4.16	7.19	11.46	13.78
	4	0.42	0.86	1.62	2.85	4.16	6.89	11.24	14.88	17.39
	8	0.86	1.63	2.98	4.15	6.90	11.48	14.87	18.53	20.47
	16	1.62	3.06	4.16	7.32	11.48	14.95	18.33	21.13	22.43
	32	3.02	4.16	6.96	11.52	15.03	18.32	21.11	22.84	23.16
	64	4.17	7.19	11.54	14.86	18.19	21.02	22.65	23.44	26.39
	128	6.98	11.51	14.87	18.20	21.03	22.61	23.74	26.52	44.91
	256	11.51	14.88	18.19	21.03	22.54	23.48	26.75	46.00	71.79
	512	14.85	18.15	20.92	22.73	23.75	26.87	45.75	75.14	107.78
	1024	16.07	19.53	21.63	23.15	26.33	45.63	72.95	104.57	150.13
	2048	17.46	20.47	22.19	25.89	45.61	73.81	104.28	153.38	191.98

more (256 thread	4096	8192
total time	9.98E-02	1.07E-01
events / sec	6.30E+07	5.89E+07
factor gpu/cpu	**212.02**	197.97

July 2020 - S. Roiser
https://docs.google.com/spreadsheets/d/1L7saADFM7Atx8hgjd-NV-IT49OLVZ-I3bA9VMZBHvbk/